molecular pharmaceutics

articles

Alignment-Free Prediction of a Drug−Target Complex Network Based on Parameters of Drug Connectivity and Protein Sequence of Receptors

Dolores Viña,[†,‡] Eugenio Uriarte,[†] Francisco Orallo,[‡] and Humberto González-Díaz*[,†,§]

Unit of Bioinformatics & Connectivity Analysis, Institute of Industrial Pharmacy, Faculty of Pharmacy, and Department of Organic Chemistry, University of Santiago de Compostela, 15782, Spain, Department of Pharmacology, University of Santiago de Compostela, 15782, Spain, and Department of Microbiology and Parasitology, University of Santiago de Compostela, 15782, Spain

Received July 22, 2008; Revised Manuscript Received January 26, 2009; Accepted March 12, 2009

Abstract: There are many drugs described with very different affinity to a large number of receptors. In this work, we selected drug−receptor pairs (DRPs) of affinity/nonaffinity drugs to similar/dissimilar receptors and we represented them as a large network, which may be used to identify drugs that can act on a receptor. Computational chemistry prediction of the biological activity based on quantitative structure−activity relationships (QSAR) substantially increases the potentialities of this kind of networks avoiding time- and resource-consuming experiments. Unfortunately, most QSAR models are unspecific or predict activity against only one receptor. To solve this problem, we developed here a multitarget QSAR (mt-QSAR) classification model. Overall model classification accuracy was 72.25% (1390/1924 compounds) in training, 72.28% (459/635) in cross-validation. Outputs of this mt-QSAR model were used as inputs to construct a network. The observed network has 1735 nodes (DRPs), 1754 edges or pairs of DRPs with similar drug−target affinity (sPDRPs), and low coverage density $d = 0.12\%$. The predicted network has 1735 DRPs, 1857 sPDRPs, and also low coverage density $d = 0.12\%$. After an edge-to-edge comparison (chi-square = 9420.3; $p < 0.005$), we have demonstrated that the predicted network is significantly similar to the one observed and both have a distribution closer to exponential than to normal.

Keywords: Drug−target complex networks; drug−receptor interaction; multitarget quantitative structure−activity relationships (mt-QSAR); molecular descriptor; Markov model; complex networks

Introduction

There is a high interest in searching rational approaches for drug discovery. In particular, *in silico* prediction is of major importance for molecular pharmaceutical sciences in this sense. The actual availability of more than 23 millions of chemistry substances and approximately 4000 new substances incorporated every day shows the impossibility of testing their biological activity on a huge number of receptors. Many lines of evidence have indicated that mathematical/computational approaches, such as structural bioinformatics,[1−5] molecular docking,[6−11] pharmacophore modeling,[12,13] protein subcellular location prediction,[14−18]

* Address correspondence to this author at Faculty of Pharmacy, University of Santiago de Compostela, 15782, Spain. Tel: +34 981563100. Fax: +34 981594912. E-mail: humberto.gonzalez@usc.es or gonzalezdiazh@yahoo.es.

[†] Unit of Bioinformatics & Connectivity Analysis, Institute of Industrial Pharmacy, Faculty of Pharmacy, and Department of Organic Chemistry.

[‡] Department of Pharmacology.

[§] Department of Microbiology and Parasitology.

(1) Chou, K. C. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* **2004**, *11*, 2105–2134.

(2) Chou, K. C. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. *Biochem. Biophys. Res. Commun.* **2004**, *316*, 636–642.

10.1021/mp800102c CCC: $40.75
Published on Web 03/12/2009

Monte Carlo simulated annealing approach,[19] diffusion-controlled reaction simulation,[20] graph/diagram approach,[21−33] biomacromolecular internal collective motion simulation,[34,35] molecular packing,[36,37] identification of membrane proteins and their types,[38] identification of enzymes and their functional classes,[39] identification of GPCR and their types,[40,41]

(3) Chou, K. C. Molecular therapeutic target for type-2 diabetes. *J. Proteome Res.* **2004**, *3*, 1284–1288.
(4) Chou, K. C. Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. *Biochem. Biophys. Res. Commun.* **2004**, *319*, 433–438.
(5) Chou, K. C. Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. *J. Proteome Res.* **2005**, *4*, 1681–1686.
(6) Chou, K. C.; Wei, D. Q.; Zhong, W. Z. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: Ibid., **2003**, *310*, 675). *Biochem. Biophys. Res. Commun.* **2003**, *308*, 148–151.
(7) Li, Y.; Wei, D. Q.; Gao, W. N.; Gao, H.; Liu, B. N.; Huang, C. J.; Xu, W. R.; Liu, D. K.; Chen, H. F.; Chou, K. C. Computational approach to drug design for oxazolidinones as antibacterial agents. *Med. Chem.* **2007**, *3*, 576–582.
(8) Wang, J. F.; Wei, D. Q.; Chen, C.; Li, Y.; Chou, K. C. Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. *Protein Pept. Lett.* **2008**, *15*, 27–32.
(9) Zhang, R.; Wei, D. Q.; Du, Q. S.; Chou, K. C. Molecular modeling studies of peptide drug candidates against SARS. *Med. Chem.* **2006**, *2*, 309–314.
(10) Gao, W. N.; Wei, D. Q.; Li, Y.; Gao, H.; Xu, W. R.; Li, A. X.; Chou, K. C. Agaritine and its derivatives are potential inhibitors against HIV proteases. *Med. Chem.* **2007**, *3*, 221–226.
(11) Zheng, H.; Wei, D. Q.; Zhang, R.; Wang, C.; Wei, H.; Chou, K. C. Screening for New Agonists against Alzheimer's Disease. *Med. Chem.* **2007**, *3*, 488–493.
(12) Sirois, S.; Wei, D. Q.; Du, Q. S.; Chou, K. C. Virtual Screening for SARS-CoV Protease Based on KZ7088 Pharmacophore Points. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1111–1122.
(13) Chou, K. C.; Wei, D. Q.; Du, Q. S.; Sirois, S.; Zhong, W. Z. Review: Progress in computational approach to drug development against SARS. *Curr. Med. Chem.* **2006**, *13*, 3263–3270.
(14) Chou, K. C.; Shen, H. B. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3*, 153–162.
(15) Chou, K. C.; Shen, H. B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1–16.
(16) Chou, K. C.; Shen, H. B. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *BBRC* **2006**, *347*, 150–157.
(17) Chou, K. C.; Shen, H. B. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **2007**, *6*, 1728–1734.
(18) Chou, K. C.; Shen, H. B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbour classifiers. *J. Proteome Res.* **2006**, *5*, 1888–1897.
(19) Chou, K. C. Energy-optimized structure of antifreeze protein and its binding mechanism. *J. Mol. Biol.* **1992**, *223*, 509–517.
(20) Chou, K. C.; Zhou, G. P. Role of the protein outside active site on the diffusion-controlled reaction of enzyme. *J. Am. Chem. Soc.* **1982**, *104*, 1409–1413.
(21) Cornish-Bowden, A. *Fundamentals of Enzyme Kinetics*; Butterworths: London, 1979; Chapter 4.
(22) Zhou, G. P.; Deng, M. H. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. *Biochem. J.* **1984**, *222*, 169–176.
(23) Myers, D.; Palmer, G. Microcomputer tools for steady-state enzyme kinetics. *Bioinformatics (original: Comput. Appl. Biosci.)* **1985**, *1*, 105–110.
(24) Chou, K. C. Graphical rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.* **1989**, *264*, 12074–12079.
(25) Kuzmic, P.; Ng, K. Y.; Heath, T. D. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. *Anal. Biochem.* **1992**, *200*, 68–73.
(26) Andraos, J. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. *Can. J. Chem.* **2008**, *86*, 342–357.
(27) Chou, K. C. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophys. Chem.* **1990**, *35*, 1–24.
(28) Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J. Biol. Chem.* **1993**, *268*, 6119–6124.
(29) Althaus, I. W.; Gonzales, A. J.; Chou, J. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. *J. Biol. Chem.* **1993**, *268*, 14875–14880.
(30) Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry* **1993**, *32*, 6548–6554.
(31) Chou, K. C.; Kezdy, F. J.; Reusser, F. Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal. Biochem.* **1994**, *221*, 217–230.
(32) Xiao, X.; Shao, S. H.; Chou, K. C. A probability cellular automaton model for hepatitis B viral infections. *Biochem. Biophys. Res. Commun.* **2006**, *342*, 605–610.
(33) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K. C. An Application of Gene Comparative Image for Predicting the Effect on Replication Ratio by HBV Virus Gene Missense Mutation. *J. Theor. Biol.* **2005**, *235*, 555–565.
(34) Chou, K. C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys. Chem.* **1988**, *30*, 3–48.
(35) Chou, K. C. Low-frequency resonance and cooperativity of hemoglobin. *Trends Biochem. Sci.* **1989**, *14*, 212.
(36) Chou, K. C.; Nemethy, G.; Scheraga, H. A. Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. *J. Am. Chem. Soc.* **1984**, *106*, 3161–3170.
(37) Chou, K. C.; Maggiora, G. M.; Nemethy, G.; Scheraga, H. A. Energetics of the structure of the four-α helix bundle in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4295–4299.
(38) Chou, K. C.; Shen, H. B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **2007**, *360*, 339–345.
(39) Shen, H. B.; Chou, K. C. EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.* **2007**, *364*, 53–59.
(40) Chou, K. C.; Elrod, D. W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.* **2002**, *1*, 429–433.
(41) Chou, K. C. Prediction of G-protein-coupled receptor classes. *J. Proteome Res.* **2005**, *4*, 1413–1418.

identification of proteases and their types,[42] protein cleavage site prediction,[43−45] and signal peptide prediction,[46,47] can timely provide very useful information and insights for both basic research and drug design, and hence, they are widely welcomed by the science community. The present study has attempted to develop a network approach, hoping that it may stimulate new strategies for drug development. Quantitative structure−activity relationship (QSAR) studies, based on drug molecular descriptors of chemical structure, may play an important role in the prediction of biological activity.[48−53] In principle, we can extend more than 1 600 different molecular descriptors to solve the former problem.[54] Our group has introduced elsewhere a Markov chain model (MCM) method named "Markov chains invariants for network simulation and design" (MARCH-INSIDE). The MARCH-INSIDE approach makes use of MCM to calculate the average values of different molecular physicochemical properties in chemical structures.[55]

Disappointingly, QSAR studies are generally based on databases which take into consideration only structurally parent compounds binding to only one single receptor.[56] In fact, there are many receptors described with very different drug susceptibility. This very high number of possible drug−receptor pairs (DRPs) may be investigated using complex networks (CNs) to regroup or cluster drugs with a similar multireceptor affinity profile. In fact, we can use CNs to study relationships between proteins, genes, RNAs, organisms, or even nonliving objects such as Web pages, but we can also develop *in silico* procedures to predict these CNs.[57−60] For instance, we predict protein−protein interactions (PINs)[61,62] or develop protein−protein structural similarity CNs (PPSS-CNs) which can be constructed by measuring the structural similarity of pairs or proteins with similar binding sites.[63] We can construct a CN of DRPs (DRP-CN) accounting for drug affinity by multiple receptors if we set a certain analogy with PPSS-CNs:

- In DRP-CN, the DRPs play the same role as the proteins in PPSS-CNs (nodes).
- In the DRP-CN, two DRPs (PDRPs) are interconnected by an edge if they have similar drug−receptor structure and, consequently, affinity (sPDRPs); in the PPSS-CNs, two proteins are interconnected if they have similar structure and, consequently, similar function.
- In the PPSS-CN, we need to measure both structure and function of each protein if we do not have a computational approach to predict them. For DRP-CN,

(42) Chou, K. C.; Shen, H. B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Commun.* **2008**, *376*, 321–325.
(43) Chou, K. C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.* **1993**, *268*, 16938–16948.
(44) Chou, K. C. Review: Prediction of HIV protease cleavage sites in proteins. *Anal. Biochem.* **1996**, *233*, 1–14.
(45) Shen, H. B.; Chou, K. C. HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. *Anal. Biochem.* **2008**, *375*, 388–390.
(46) Chou, K. C.; Shen, H. B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **2007**, *357*, 633–640.
(47) Shen, H. B.; Chou, K. C. Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem. Biophys. Res. Commun.* **2007**, *363*, 297–303.
(48) Du, Q. S.; Mezey, P. G.; Chou, K. C. Heuristic Molecular Lipophilicity Potential (HMLP): A 2D-QSAR Study to LADH of Molecular Family Pyrazole and Derivatives. *J. Comput. Chem.* **2005**, *26*, 461–470.
(49) Du, Q. S.; Huang, R. B.; Wei, Y. T.; Du, L. Q.; Chou, K. C. Multiple Field Three Dimensional Quantitative Structure-Activity Relationship (MF-3D-QSAR). *J. Comput. Chem.* **2008**, *29*, 211–219.
(50) Du, Q. S.; Huang, R. B.; Chou, K. C. Review: Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. *Curr. Protein Pept. Sci.* **2008**, *9*, 248–259.
(51) Du, Q. S.; Huang, R. B.; Wei, Y. T.; Pang, Z. W.; Du, L. Q.; Chou, K. C. Fragment-Based Quantitative Structure-Activity Relationship (FB-QSAR) for Fragment-Based Drug Design. *J. Comput. Chem.* **2009**, *30*, 295–304.
(52) Prado-Prado, F. J.; Gonzalez-Diaz, H.; de la Vega, O. M.; Ubeira, F. M.; Chou, K. C. Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorg. Med. Chem.* **2008**, *16*, 5871–5880.
(53) Dea-Ayuela, M. A.; Perez-Castillo, Y.; Meneses-Marcel, A.; Ubeira, F. M.; Bolas-Fernandez, F.; Chou, K. C.; Gonzalez-Diaz, H. HP-Lattice QSAR for dynein proteins: Experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence. *Bioorg. Med. Chem.* **2008**, *16*, 7770–7776.
(54) Todeschini, R. and Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: New York, 2002; pp 1−520.
(55) González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Medicinal Chemistry and Bioinformatics - Current Trends in Drugs Discovery with Networks Topological Indices. *Curr. Top. Med. Chem.* **2007**, *7*, 1025–1039.
(56) Kubinyi, H. Quantitative structure-activity relationships (QSAR) and molecular modelling in cancer research. *J. Cancer Res. Clin. Oncol.* **1990**, *116*, 529–537.
(57) Bornholdt, S. and Schuster, H. G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*; WILEY-VCH GmbH & CO. KGa.: Wheinheim, 2003; pp 1−394.
(58) Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. U. Complex networks: Structure and dynamics. *Phys. Rep.* **2006**, *424*, 175–308.
(59) Réka, A.; Barabasi, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47–97.
(60) Gonzalez-Diaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Medicinal chemistry and bioinformatics - current trends in drugs discovery with networks topological indices. *Curr. Top. Med. Chem.* **2007**, *10*, 1015–1029.
(61) Chou, K. C.; Cai, Y. D. Predicting protein-protein interactions from sequences in a hybridization space. *J. Proteome Res.* **2006**, *5*, 316–322.
(62) Shi, T. L.; Li, Y. X.; Cai, Y. D.; Chou, K. C. Computational methods for protein-protein interaction and their application. *Curr. Protein Pept. Sci.* **2005**, *6*, 443–449.
(63) Zhang, Z.; Grigorov, M. G. Similarity networks of protein binding sites. *Proteins* **2006**, *62*, 470–478.

we need to measure the affinity of the drug on different receptors (DRPs affinity) if we cannot predict it.

We propose herein, for the first time, to reconstruct a DRP-CN taking into consideration only the sequence of the receptor and the chemical connectivity of the drug, but without relying on geometry optimization or drug−drug and target−target alignment. The task is difficult, but interesting, because we pretend to shun using 3D structures of drug, receptor, and drug−receptor complex (Docking) as well as drug−drug superposition (CoMFA methods) or sequence−sequence alignment for receptors.[64,65] A method independent of these aspects may become notably faster because we do not have to run optimization algorithms to predict the 3D structure; these optimization algorithms are computationally expensive and not fully accurate for many proteins and/or protein−drug complexes. The present alignment-free method is also significant because an alignment method may fail if there are no similar function-annotated sequences in the database;[66] conversely, the QSAR approach to CNs circumvents the alignment by using structural parameters.[67] Consequently, if we set out to use the QSAR method to construct the PDRP-CN, we have to develop a QSAR able to predict DRPs. First, we developed the DRPs-mtQSAR (multitarget-QSAR)[68−71] classification model, and subsequently, we used the model outputs to construct a DRP-CN. The QSAR model proposed here is the first able to discriminate between two DRPs (PDRPs) that have similar/dissimilar affinity (sPDRPs/nPDRPs). Lastly, we compared the DRP-CN predicted with a DRP-CN constructed here, based on measured values of DRP affinity.

Methods

Computational Methods. The MARCH-INSIDE approach is based on the calculation of different physicochemical molecular properties (λ_d) as an average of atomic properties (λ_j). For instance, it is possible to derive average estimations of electrostatic or van der Waals potentials, as well as the molecular electronegativities (χ_d), refractivities (MR_d), polarizabilities (α_d), logarithms of water/*n*-octanol partition coefficients ($\log P_d$), and hardness (η_d) that we are going to use in this work, as seen in eq 1:[72]

$$\lambda_d = \frac{1}{6}\sum_{k=0}^{5} {}^k\lambda = \frac{1}{6}\sum_{k=0}^{5}\sum_j p_k(\lambda_j)\cdot\lambda_j \quad (1)$$

It is possible to consider isolated atoms ($k = 0$) in the estimation of the molecular properties ${}^0\eta$, ${}^0\chi$, ^{0}MR, ${}^0\alpha$, $\log {}^0P$. In this case, the probabilities ${}^0p(\lambda_j)$ are determined without considering the formation of chemical bonds (simple additive scheme). However, it is possible to consider the gradual effects of the neighboring atoms at different distances in the molecular backbone. In order to reach this goal, the method uses a MM that determines the absolute probabilities ${}^kp(\lambda_j)$ with which the atoms placed at different distances k affect the contribution of the atom j to the molecular property in question.

$$k_\lambda = [{}^0p(\lambda_1){}^0p(\lambda_2)\ldots{}^0p(\lambda_n)]\cdot\begin{bmatrix} 1_{p_{1,2}} & 1_{p_{1,2}} & 1_{p_{1,3}} & \cdot & 1_{p_{1,n}} \\ 1_{p_{2,1}} & 1_{p_{2,2}} & 1_{p_{2,3}} & \cdot & 1_{p_{2,n}} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1_{p_{n,1}} & \cdot & \cdot & \cdot & 1_{p_{n,n}} \end{bmatrix}^k \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdot \\ \cdot \\ \lambda_n \end{bmatrix} = \sum_{j=1}^{n} {}^kp(\lambda_j)\cdot\lambda_j \quad (2)$$

where, from left to right, the first term is ${}^k\lambda$, which is the average molecular property considering the effects of all the atoms placed at distance k over every atomic property λ_j. The vector on the left-hand side of the equation contains the probabilities ${}^0p(\lambda_j)$ of every atom in the molecule, without considering any chemical bonds. The matrix in the center of the equation is the so-called stochastic matrix. The values of this matrix (${}^1p_{ij}$) are the probabilities with which each atom affects the parameters of the atom bonded to it. Both kinds of probabilities ${}^0p(\lambda_j)$ and ${}^1p_{ij}$ are easily calculated from the atomic parameters (λ_j) and the chemical bonding information:

$${}^0p_{ij} = \frac{\lambda_j}{\sum_{k=1}^{n}\lambda_k} \quad (3)$$

(64) Barbany, M.; Gutierrez-de-Teran, H.; Sanz, F.; Villa-Freixa, J. Towards a MIP-based alignment and docking in computer-aided drug design. *Proteins* **2004**, *56*, 585–594.
(65) Cramer, I., R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(66) Dobson, P. D.; Doig, A. J. Predicting enzyme class from protein structure without alignments. *J. Mol. Biol.* **2005**, *345*, 187–199.
(67) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750–778.
(68) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between drugs and nondrugs by prediction of activity spectra for substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432–2437.
(69) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of activity spectra for biologically active substances. *Bioinformatics* **2000**, *16*, 747–748.
(70) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.
(71) Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M. N. D. S.; Cagide Fajin, J. L.; Morell, C.; Molina Ruiz, R.; Cañizares-Carmenate, Y.; Dominguez, E. R. *J. Comb. Chem.* **2008**, *10*, 897–913.
(72) Santana, L.; Uriarte, E.; González-Díaz, H.; Zagotto, G.; Soto-Otero, R.; Mendez-Alvarez, E. A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *J. Med. Chem.* **2006**, *49*, 1149–56.

$$ {}^{1}p_{ij} = \frac{\delta_{ij}\cdot\lambda_j}{\sum_{k=1}^{n}\delta_{ik}\cdot\lambda_k} \quad (4) $$

The only difference is that in the probabilities ${}^{0}p(\lambda_j)$ we consider isolated atoms by carrying out the sum in the denominator over all n atoms in the molecule. On the other hand, for ${}^{1}p_{ij}$ the chemical bonding is taken into consideration by means of the factor δ_{ij}. This factor has the value 1 if atoms i and j are chemically bonded and it is 0 otherwise. All calculations were performed using the program MARCH-INSIDE version 3.0,[73] which can be obtained for free academic use, upon request, from the corresponding author of the present work.

In the same way, different physicochemical molecular properties of the receptor were calculated splitting the receptor in six different regions (f_x) beginning in x (x = 0, 61, 121, 181, 241, 301aa) with 60 amino acids each, except the last one, and taking into account the amino acid composition. Therefore, the properties in the sequence ($\lambda_r(f_x)$) were determined as the property-sum of each amino acid (λ_{aj}) multiplied by the number of this amino acid (N_{aj}) divided between the total of amino acids (N_f) in the corresponding fragment of receptor (f_x). In Tables 1SM and 2SM in the Supporting Information, we depict the values of λ_j and λ_{aj} for some atoms, amino acids and receptors respectively. All these values depend only on the atomic standard parameters (λ_j) obtained from the literature.[54,74]

$$ \lambda_r(f_x) = \sum_{aj} p(\lambda_{aj})\cdot\lambda_{aj} = \sum_{aj}\frac{N_{aj}(f_x)}{N_f}\cdot\lambda_{aj} = \sum_{aj}\left(\frac{N_{aj}(f_x)}{N_f}\right)\cdot\left(\sum_{j}\frac{\lambda_j}{N_j}\right) \quad (5) $$

Statistical Analysis. Let λ_d be drug molecular descriptors and $\Delta\lambda(f_x, d)$ drug−receptor interaction descriptors for different drugs (d) with different receptor fragment sequences (f_x); we attempt to develop a simple linear classifier of mt-QSAR type with the general formula

$$ S_{pred} = \sum_{\lambda=0}^{4} b(\lambda)\cdot\lambda_d + \sum_{f=0}^{5}\sum_{\lambda=0}^{4} b(f_x,\lambda)\cdot\Delta\lambda(f_x,d) + b = \sum_{\lambda=0}^{4} b(\lambda)\cdot\lambda_d + \sum_{f=0}^{5}\sum_{\lambda=0}^{4} b(f_x,\lambda)\cdot(\lambda_r(f_x) - \lambda_d) + b \quad (6) $$

We used linear discriminating analysis (LDA) to fit this discriminant. The model deals with the classification of a compound set with or without affinity to different receptors. A dummy variable affinity class (AC) was used as input to codify the affinity. This variable indicates either high (AC = 1) or low (AC = 0) affinity of the drug to the receptor. S_{pred} (affinity predicted score) is the output of the model, and it is a continuous adimensional score that sorts compounds from low to high affinity. In eq 6, b represents the coefficients of the classification function, determined by the LDA module of the STATISTICA 6.0 software package.[75] We used forward stepwise algorithm for a variable selection. The statistical significance of the LDA model was determined calculating the p-level (p) of error with chi-square test. We also inspected the specificity, sensitivity, and total accuracy to determine the quality-of-fit to data in training. The validation of the model was corroborated with external prediction series.

Data Set. The data set was conformed to a set of marketed and/or reported drugs/receptor pairs where affinity/nonaffinity of drugs with the receptors was established taking into consideration the IC_{50}, k_i, pk_i,... values. In consequence, we managed to collect 1735 cases (DRPs). In addition, we used a negative control series of DRPs conformed to real drugs and chimera receptors (824 cases). The sequence of the chimera receptors was built up assembling random fragments of different real receptors. In the two data sets used, there were the following training series: 454 compounds showing affinity to the receptor plus 1470 compounds with nonaffinity (1924 in total). Predicting series: 149 + 486 = 635 in total. The names or codes for all compounds are depicted in Table 3SM in the Supporting Information, due to space constraints, as well as the references consulted to compile the data in this table.

Drug−Receptor Pair (DRP) Complex Network (CN) Construction. In order to achieve the drug−receptor affinity with a network approach where one node represents a DRP and the edges show similarity between two nodes related to the activity (sPDRPs or nPDRPs), we carried out the following steps:

1. First, we calculated two types of affinity Z-scores (drug score and receptor score) for both experimental and QSAR-predicted values:

$$ z_{obs}(d) = \frac{o\cdot(s_{obs}(d,r) - mean\cdot s_{obs}(d,r))}{1 + SD\cdot s_{obs}(d,r)} \quad (7a) $$

$$ z_{pred}(d) = \frac{(s_{pred}(d,r) - mean\cdot s_{pred}(d,r))}{1 + SD\cdot s_{pred}(d,r)} \quad (7b) $$

$$ z_{obs}(r) = \frac{o\cdot(\text{mean}\cdot s_{obs}(d,r) - 1)}{1 + SD\cdot s_{obs}(d,r)} \quad (8a) $$

$$ z_{pred}(r) = \frac{(\text{mean}\cdot s_{pred}(d,r) - 1)}{1 + SD\cdot s_{pred}(d,r)} \quad (8b) $$

where s is the score affinity, either observed score (s_{obs}) or predicted score (s_{pred}). s_{obs} was calculated on the experimental

(73) González-Díaz, H., Molina-Ruiz, R. and Hernandez, I. *MARCH-INSIDE* v3.0 (Markov Chains Invariants for Simulation & Design); Windows supported version under request to the main author, contact e-mail: gonzalezdiazh@yahoo.es, 3.0; 2007.

(74) Hou, T.; Xu, X. ADME Evaluation in Drug Discovery. 2. Prediction of Partition Coefficient by Atom-additive Approach Based on Atom-weighted Solvent Accessible Surface Areas. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1058–1067.

(75) StatSoft. Inc. *STATISTICA* (data analysis software system), version 6.0, www.statsoft.com.Statsoft, Inc., 6.0; 2002.

data (IC_{50}, k_i, pk_i,...). We calculated the s_{pred} of each one of the 1735 drugs with all the receptors studied here by substituting the molecular descriptors into the QSAR equation, using the Microsoft Excel application.[76] Mean is the average either of s_{obs} or s_{pred} for the DRP. SD is the standard deviation, and o is a variable with values 1 (when s_{obs} is proportional to the biological affinity) and −1 (when s_{obs} is opposite to the biological affinity).

2. We calculated the distance matrix between all PDRP using a Euclidean distance:

$$^{obs}d_{ij} = \sqrt{(z_{obs}(d)_i - z_{obs}(d)_j)^2 + (z_{obs}(r)_i - z_{obs}(r)_j)^2} \quad (9a)$$

$$^{pred}d_{ij} = \sqrt{(z_{pred}(d)_i - z_{pred}(d)_j)^2 + (z_{pred}(r)_i - z_{pred}(r)_j)^2} \quad (9b)$$

3. Using Microsoft Excel[76] again, we transformed the DRPs distance matrices into Boolean matrices. The elements of this type of matrix are equal to 1 if a PDRPs has a Euclidean distance d_{ij} < a cutoff value. We explored the threshold values in a range from 0.002 to 0.05 trying to obtain an average DRP node degree equal to 3 and minimizing the number of disconnected DRPs. The line command used in Excel to transform the distance matrix into a Boolean matrix was f = if (A$1= $B2, 0, if (B2 > cut-off, 0, 1)). It allows transforming distance into Boolean values and equals the main diagonal elements to 0, avoiding loops in the future network.[77]

4. We compared the observed and predicted PDRP pair-to-pair networks using a chi-square (χ^2) test. Therefore, we used a contingency table where a, b, c and d are the observed frequencies in our networks. (See Table 1.) These frequencies were calculated as follows: f = if (and (obs B2! = 1, pred B2! = 1), 1, if (and (obs B2! = 1, pred B2! = 0), 2, if (and (obs B2! = 0, pred B2! = 1), 3, 4)). Then

- "a" is the number of PDRPs neither connected in observed networks nor in predicted ones (the elements in the observed and predicted matrices are equal to 0);
- "b" is the number of PDRPs not connected in observed networks but connected in predicted ones (observed is 0 and predicted is 1);
- "c" is the number of PDRPs connected in observed networks but not connected in predicted ones (observed is 1 and predicted is 0);
- "d" is the number of PDRPs connected in observed networks and in predicted ones (observed and predicted are 1).

5. The chi-square test allows us to determine if the variables are associated or not. If they are not associated, we could conclude that they are independent. The first step of the chi-square test for independence is to establish hypotheses. A null hypothesis occurs when the two variables are independent (the observed and predicted activity of the DRPs is not associated). The alternative hypothesis to be tested occurs when the two variables are dependent. χ^2 was calculated as follows:[78]

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (10)$$

6. In eq 10 O_{ij} is the observed frequency and E_{ij} is the expected or theoretical frequency. E_{ij} is calculated as follows:

$$E_{11} = \frac{(a+b)\times(a+c)}{n} \quad (11a)$$

$$E_{21} = \frac{(c+d)\times(a+c)}{n} \quad (11b)$$

$$E_{12} = \frac{(a+b)\times(b+d)}{n} \quad (11c)$$

$$E_{22} = \frac{(c+d)\times(b+d)}{n} \quad (11d)$$

7. Then, we compared the value calculated in the formula above to a standard set of tables. The value returned from

Table 1. Contingency Table Results for QSAR and CN Analyses

parameters			observed values[a]		
parameter	%	predicted	iDRPs	aDRPs	total
Mt-QSAR Model Training					
specificity	72.7	iDRPs	1 069	401	1 470
sensitivity	70.7	aDRPs	133	321	454
accuracy	72.2	total	1 202	722	1 924
mt-QSAR Model Cross-Validation					
specificity	72.4	iDRPs	352	134	486
sensitivity	71.8	aDRPs	42	107	149
accuracy	72.2	total	394	241	635
QSAR-χ^2	376.19				
p	<0.001				

Observed vs Predicted Complex Networks					
parameter	%	predicted	nPDRPs	sPDRPs	total
nonsimilar PDRPs	99.9	nPDRPs	3 001 474	3 302	3 004 776
similar PDRPs	5.5	sPDRPs	3 508	206	3 714
accuracy	99.8	total	3 004 982	5 243	3 010 225
Expected Contingency for Complex Networks[b]					
nonsimilar PDRPs	99.8	nPDRPs	2 999 542.5	5 233.5	3 004 776
similar PDRPs	0.2	sPDRPs	5 439.5	9.5	5 449
accuracy	99.6	total	3 004 982	5 243	3 010 225
CN-χ^2	9 420.3				
p	<0.005				

[a] aDRPs: drug−receptor pairs for compounds with high affinity. iDRPs: drug−receptor pair for compounds with low affinity. sPDRPs: pairs of similar drug−receptor pairs. nPDRPs: no-similar pairs of drug−receptor pairs. [b] Expected values.

(76) Microsoft Corp. *Microsoft Excel 2002*.

(77) González-Díaz, H.; Prado-Prado, F. Unified QSAR and Network-Based Computational Chemistry Approach to Antimicrobials, Part 1: Multispecies Activity Models for Antifungals. *J. Comput. Chem.* **2008**, *29*, 656–657.

(78) Hill, T.; Lewicki, P. *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*; StatSoft: Tulsa, 2006; pp 1−813.

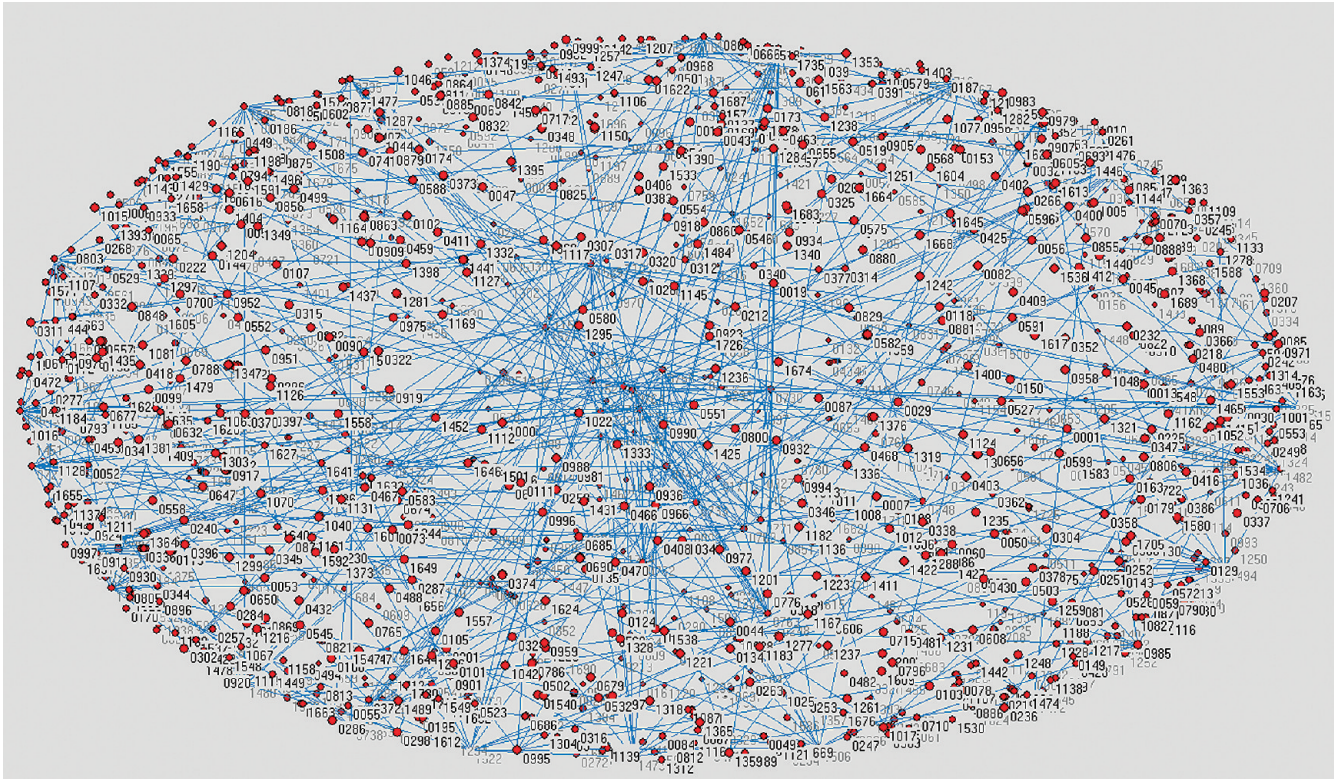

Figure 1. Graphical view of the coincident edges on observed vs predicted DRP-CNs.

the table is $p < 0.005$. Thus, we can reject the null hypothesis and conclude that there is an association between the variables.

8. The Boolean matrix was saved as a.txt format file. After we had renamed the.txt file as a.mat file we read it with the CentiBin software.[79,80] Using CentiBin we can not only represent the network but also highlight all DRP (nodes) connected to a specific DRP and consequently calculating connectivity parameters, including the node degree.

9. CentiBin software was used to generate random networks by five different algorithms including Barabasi−Albert random network, Kleinberg small world network (SWN), 2D lattice network, Erdos−Renyi network and Epsstein power law network (PLN).[80] These random networks were compared with the observed and predicted networks.

10. Lastly, all node degrees were used as input in STATISTICA in order to study the distribution of the network and compare it with other ideal network distributions including normal, log-normal, exponential, gamma, and chi-square.[75]

Results

Training and Validation of the mt-QSAR Model. Common physicochemical properties have been demonstrated to be useful on protein QSAR.[81,82] This work introduces for the first time a single linear mt-QSAR equation model to classify drugs according to their affinity to bound more than 60 different receptors based on simple physicochemical parameters. The best model found was

$$S_{\text{pred}} = 85.4 \cdot \alpha(d) - 70.5 \cdot \alpha(f_{0aa}, d) - 0.5 \cdot \alpha(f_{121aa}, d) - 15.0 \cdot \alpha(f_{181aa}, d) - 51.0 \cdot \log P(d) + 51.0 \cdot \log P(f_{61aa}, d) + 2.9 \cdot \chi(f_{241aa}, d) + 20.5 \cdot \eta(d) - 15.8 \cdot \eta(f_{0aa}, d) - 7.1 \cdot \eta(f_{181aa}, d) - 1.3N = 1924 \qquad \chi^2 = 376.19 \qquad p < 0.001 \quad (12)$$

where α_d is the polarizability, $\log P_d$ is the logarithm of the water/*n*-octanol partition coefficient, η is the hardness, χ_d is the molecular electronegativity, N is the number of cases (DRPs) used to train the model and chi-square (χ^2) is the statistic used to demonstrate that the model significantly discriminates between DRPs of compounds with affinity (aDRP) or nonaffinity (iDRP) to the receptor, at a $p < 0.001$ level of error. Following the notation given above, for example $\alpha(f_{0bp},d) = \alpha(d) - \alpha(f_{0bp})$ is the difference between the polarizability of the drug and the receptor region from 0 to 60bp. In Table 1SM in the Supporting Information, we

(79) Koschützki, D. *CentiBiN* Version 1.4.2, 2006.

(80) Junker, B. H.; Koschutzki, D.; Schreiber, F. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* **2006**, *7*, 219.

(81) Ivanciuc, O.; Oezguen, N.; Mathura, V. S.; Schein, C. H.; Xu, Y.; Braun, W. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr. Med. Chem.* **2004**, *11*, 583–593.

(82) Schein, C. H.; Ivanciuc, O.; Braun, W. Common physical-chemical properties correlate with similar structure of the IgE epitopes of peanut allergens. *J. Agric. Food Chem.* **2005**, *53*, 8752–9.

depict all the receptor parameters necessary to evaluate a compound with the QSAR. For a new receptor, it would be necessary to calculate the new parameters using Table 2SM in the Supporting Information. This model, with ten variables, classifies correctly 321 out of 454 aDRP (drug−receptor pairs for compounds with high affinity) (sensitivity of 70.70%) and 1069 out of 1470 iDRP (drug−receptor pair for compounds with low affinity) (specificity of 72.72%). Overall training accuracy was 72.25% (1390 out of 1924 DRPs). The validation of the model was carried out by means of external predicting series. The model classifies correctly 107 out of 149 aDRP (71.81%) and 352 out of 486 iDRP (72.42%) in validation series. Accuracy for validation series (predictability) was 72.28% (459 out of 635 DRPs). These results (Table 1) indicate that we developed an accurate model according to previous reports on the use of LDA in QSAR.[83,84]

Complex Network Study. In order to recall the capacity of the mt-QSAR to predict new CNs we selected a database of recently assayed drugs instead of using the same data employed for the DT-CN. With these goals in mind, we constructed first a new observed DRP-CN, considering the experimental data and exploring the threshold values in a range from 0.0002 to 0.05, obtaining an average degree from 1.91 to 27.65 respectively (see Table 2). Finally, a cutoff = 0.002 was selected to obtain average DRP node degree equal to 3.02 and 0 was the number of disconnected DRPs. Node degree equal to 3 was selected because if it was 1, failure probability would be too high for the model and more than 3 would be difficult to compare. Next, we used the mt-QSAR equation to predict the scores of biological affinity for 1735 DRPs including all the 60 receptors studied. The same as before, we explored the threshold values in a range from 0.002 to 0.05, obtaining an average degree from 1.13 to 20.34 respectively, a cutoff = 0.007, which leads to an average degree of 3.14, 0 was selected as the number of disconnected DRPs. Additionally, with this threshold, there were 1754 edges for the observed network and 1857 for the predicted network.

In Figure 1, we illustrate visually the complex relationships between DRPs, drawing coincident edges for both the observed and predicted DRP-CNs. The numeric labels of the nodes identify the different inputs (DRPs) used in the analysis. In order to compare the observed and predicted networks, we used a chi-square test; the obtained value for the $p < 0.005$ error level was chi-square = 9420.3.

Table 2. Elemental Network Properties for Different Cutoff Values

	real network			predicted network		
cutoff	av degree	disc. nodes	edges	av degree	disc. nodes	edges
0.0002	1.91	0	793	1.13	0	114
0.0004	2.08	0	940	1.18	0	157
0.0006	2.21	0	1053	1.23	0	200
0.0008	2.33	0	1150	1.31	0	268
0.0010	2.42	0	1236	1.36	0	309
0.0020	3.02	0	1754	1.71	0	618
0.0030	3.51	0	2176	2.03	0	890
0.0040	4.13	0	2711	2.32	0	1149
0.0050	4.55	0	3078	2.57	0	1366
0.0060	5.12	0	3570	2.88	0	1630
0.0070	5.62	0	4012	3.14	0	1857
0.0080	6.02	0	4353	3.37	0	2056
0.0090	6.48	0	4750	3.86	0	2278
0.0100	7.12	0	5366	3.86	0	2481
0.0200	12.72	0	10165	6.86	0	5081
0.0250	14.94	0	12093	8.48	0	6492
0.0300	17.26	0	14102	10.66	0	8384
0.0350	19.44	0	15994	12.94	0	10361
0.0400	21.63	0	17900	15.35	0	12452
0.0450	24.13	0	20066	17.57	0	14374
0.0500	27.65	0	23121	20.34	0	16775

Discussion

The mt-QSAR Model for DRPs. To the best of our knowledge, the present is the first mt-QSAR model with the probability of binding organic compounds to very large diversity of receptors based only on the molecular connectivity of the drug and the sequence of receptor fragments. Two possible applications for the present model are the biomolecular screening of drug affinity to different receptors and the construction of multireceptor affinity profile networks for drugs. In both cases, receptor susceptibility identification is imperative. In order to use the present DRPs-mtQSAR model for new compounds and/or receptors, first we have to calculate the physicochemical parameters of the new compounds and/or receptors. Next, we can substitute these values in the eq 12 to predict the DRPs. Subsequently, we can use the model outputs to construct a DRP-CN assigning new nodes and edges (sPDRPs/nPDRPs) for the new cases. The final prediction bases on both the values predicted by the QSAR and the correct connection of the new nodes to former nodes with similar drugs and/or receptors. In Supporting Information (Table 4SM) we give many examples of drug and receptor cases with detailed information on the compounds, predicted classification, and probability of affinity to different receptors of the drugs.

Using this model we can construct QSAR-based charts to depict visually the different relationships between the drug−receptor affinity score (analysis of desirability) and different regions of the receptors for the same physicochem-

(83) Alvarez-Ginarte, Y. M.; Marrero-Ponce, Y.; Ruiz-Garcia, J. A.; Montero-Cabrera, L. A.; Vega, J. M.; Noheda Marin, P.; Crespo-Otero, R.; Zaragoza, F. T.; Garcia-Domenech, R. Applying pattern recognition methods plus quantum and physico-chemical molecular descriptors to analyze the anabolic activity of structurally diverse steroids. *J. Comput. Chem.* **2007**, *29*, 317–333.

(84) Morales, A. H.; Rodríguez-Borges, J. E.; García-Mera, X.; Fernández, F.; Dias-Sueiro-Cordeiro, M. N. Probing the Anticancer Activity of Nucleoside Analogues: A QSAR Model Approach Using an Internally Consistent Training Set. *J. Med. Chem.* **2007**, *50*, 1537–1545.

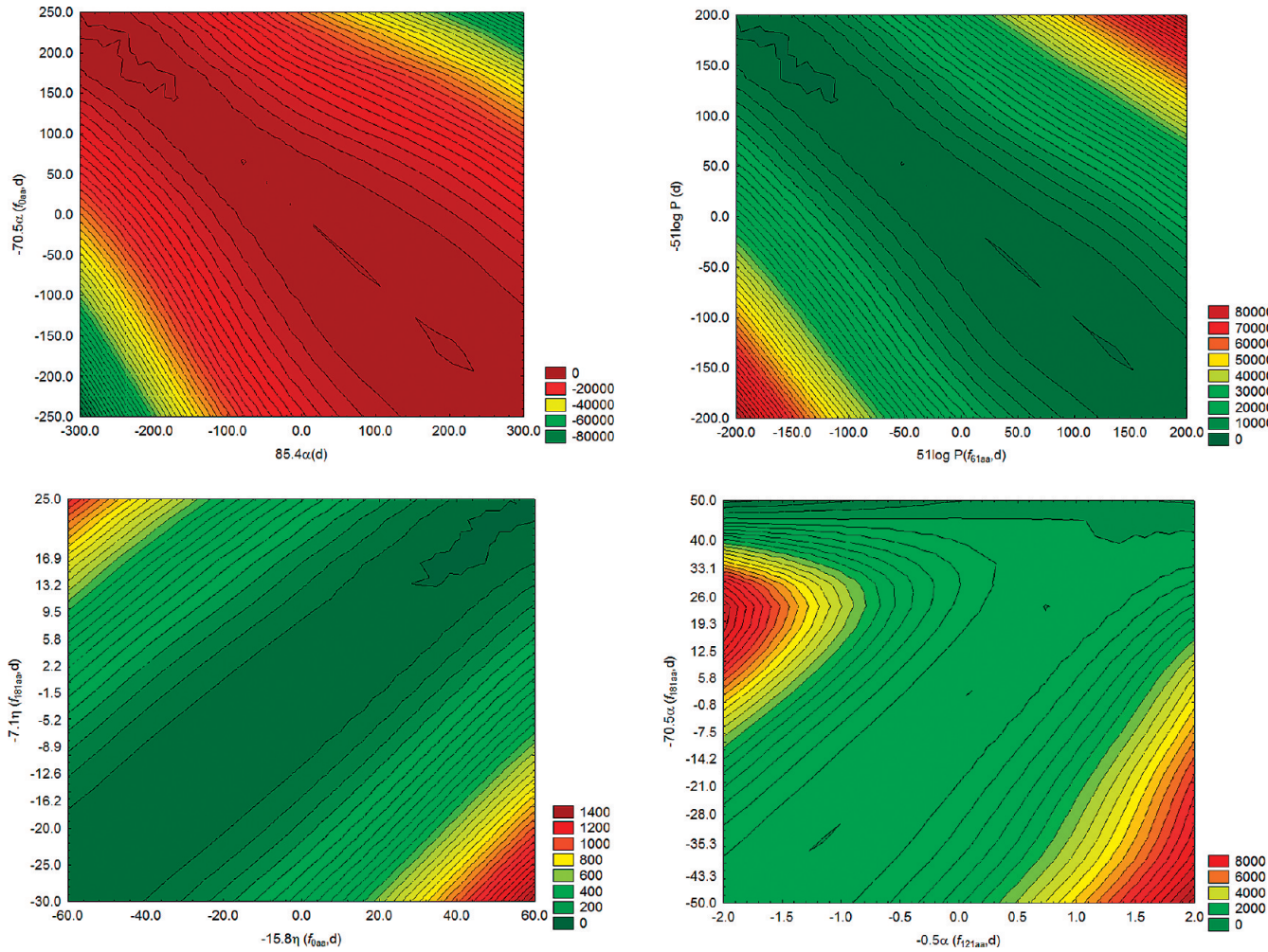

Figure 2. Example of chart used for the desirability analysis of the QSAR model.

ical property.[85] It could be used to optimize the drug or the receptor, changing only one property by organic synthesis modification of drug or genetic engineering of the receptor. In Figure 2, we illustrate some of these charts. Note that these charts may refer to only one receptor region or two different regions at the same time.

Comparison of the Observed vs Predicted DRP Complex Networks. Molecular CNs are used to study large databases and/or complex systems.[86−88] Proteins, nucleic acids, and small molecules form a dense network of molecular interactions in a cell.[89] A DRP-CN may help us to detect the most similar cases (sPDRPs) in the observed database and study the complexity of drug−target interaction phenomena. We can conclude, based on the chi-square test presented above, in the Results section, that there is a statistically significant similarity between the observed and predicted networks when the average node degree is above 3. We also conclude that even when both (observed and predicted networks) node distributions do not fit significantly (neither to exponential nor to normal), both have a distribution closer to exponential than to normal (Figure 3). In Table 1 we illustrate the contingency matrix with the number of sPDRP (similar DRPs or same number of DRP-to-DRP connections) and nPDRP (nonsimilar DRPs) for both CNs (observed and predicted networks). Notably, in the contingency table of the observed vs predicted CNs, we find both a high number of nPDRPs and a low number of sPDRPs, coinciding with the expected behavior when accuracy values are higher than 99%. It indicates that the predicted DRP-CN reproduces correctly the high specificity of the drugs investigated by the receptors.

Interconnection of the mt-QSAR Based DRP-CNs with Other CNs. Barabasi et al.[90] have very recently reported the construction of a drug−target CN (DT-CN) based on US Food and Drug Administration-approved drugs

(85) González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. A Model for the Recognition of Protein Kinases Based on the Entropy of 3D van der Waals Interactions. *J. Proteome Res.* **2007**, *6*, 904–908.

(86) Bonchev, D.; Buck, G. A. From molecular to biological structure and back. *J. Chem. Inf. Model.* **2007**, *47*, 909–917.

(87) Bonchev, D. On the complexity of directed biological networks. *SAR QSAR Environ. Res.* **2003**, *14*, 199–214.

(88) Park, J.; Barabasi, A. L. Distribution of node characteristics in complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17916–17920.

(89) Spirin, V.; Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12123–12128.

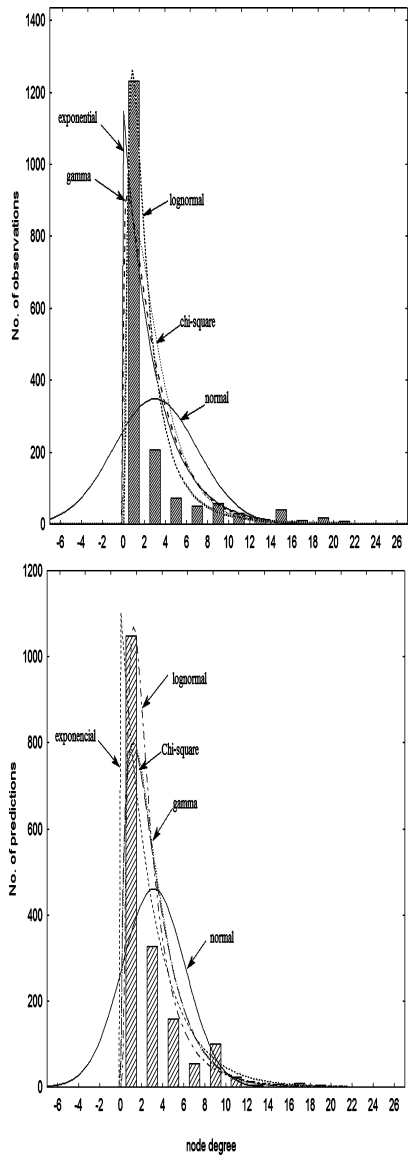

Figure 3. Node degree distributions for both observed and predicted DRP-CNs.

and proteins linked by drug−target binary associations. The DT-CN is bipartite by definition because they set two types of nodes: drug nodes or protein receptor node (drug targets). We propose here a different representation for the same problem because the DRP-CN is not bipartite, but it considers only one type of node. As referred above, each node represents here both a drug and a receptor (DRPs). In this sense, the nodes of our CN are the edges of the DT-CN, so the present DRP-CN can be defined as the first line graph (LG) of the DT-CN.[91] In addition, Yamanishi et al.[92] also reported a predictive algorithm to construct DRPs-CN. In this article, the authors characterized four classes of DRP-CNs in humans involving enzymes, ion channels, G-protein-coupled receptors (GPCRs) and nuclear receptors, and revealed significant correlations between drug structure similarity, target sequence similarity and the drug−target interaction network topology. The originality of the results, according to Yamanishi et al.,[92] lies in the formalization of the DRP interaction inference as a supervised learning problem for a bipartite graph, the lack of need for 3D structure information of the target proteins, and in the integration of chemical and genomic spaces into a unified space that they called "pharmacological space". However, the method relies upon both drug−drug graph alignment and receptor−receptor sequence alignment. The drawbacks of the use of alignment-dependent methods in this kind of work have been pointed out by Dobson and Doig[66,93] and reviewed in detail by Han et al.[94] Alignment fails when a similar protein cannot be identified, or when any similar proteins identified also lack reliable annotations. In this sense, the importance of the present study does not rely, undoubtedly, on the construction of the LG version of the Barabasi's DT-CN. The importance of our study is that we can construct the CN starting from experimental outcomes as in the work of Barabasi et al.[90] and predict the CN for new DRPs not experimentally determined using the mt-QSAR as in the work of Yamanishi et al., but we do not have to rely upon drug−drug or target−target (receptor−receptor) alignment. It allows us to add computationally new DRPs describing new potential drugs or targets that had not been described before in an alignment-independent way. We selected here the LG graph only to facilitate the construction of the CN from the outputs of the mt-QSAR model. In closing, the DRP-CNs predicted with this procedure could become a useful tool to identify potential drugs and/or targets. Both drugs and/or targets could be incorporated to DT-CNs; or the new targets added to a human disease−target CN[95] or the new drugs included into a disease−drug CN.[96]

Conclusions

Using the MARCH-INSIDE approach, it is possible to seek a mt-QSAR classifier to predict the probability of drugs to bind more than 60 different molecular receptors based only on drug connectivity and receptor sequence. The model

(90) Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25*, 1119–26.

(91) Estrada, E.; Guevara, N.; Gutman, I. Extension of Edge Connectivity Index. Relationships to Line Graph Indices and QSPR Applications. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 428–431.

(92) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, i232−40.

(93) Dobson, P. D.; Cai, Y. D.; Stapley, B. J.; Doig, A. J. Prediction of protein function in the absence of significant sequence similarity. *Curr. Med. Chem.* **2004**, *11*, 2135–2142.

(94) Han, L.; Cui, J.; Lin, H.; Ji, Z.; Cao, Z.; Li, Y.; Chen, Y. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* **2006**, *6*, 4023–4037.

(95) Goh, K. I.; Cusick, M. E.; Valle, D.; Childs, B.; Vidal, M.; Barabasi, A. L. The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 8685–8690.

(96) Nacher, J. C.; Schwartz, J. M. A global view of drug-therapy interactions. *BMC Pharmacol.* **2008**, *8*, 5.

can be used as a tool for preliminary screening of drugs without relying upon geometrical optimization of drug, receptor, and drug−receptor complex structure and avoiding receptor alignment as well. This mt-QSAR was also demonstrated to be an efficient tool for computational assembly of drug−target complex networks that accurately reproduces the network based on experimental findings. This kind of complex network could become a valuable approach to explore large drug−target data of high complexity.

Acknowledgment. H.G.-D. and D.V. acknowledge sponsorships for a tenure-track research position at the University of Santiago de Compostela from the *Isidro Parga Pondal* Programme of the "Dirección Xeral de Investigación e Desenvolvemento, Xunta de Galicia". F.O. acknowledges partial financial support of Programme for Promotion of Research Activity, SUG 2007-2008, from "Consellería de Educación e Ordenación Universitaria de la Xunta de Galicia". We are grateful to the "Direccion Xeral de I+D+I, Xunta de Galicia", Project INCITE08PXIB203022PR, for the financial support.

Supporting Information Available: Table 1SM [receptor properties that enter into the QSAR equation], Table 2SM [atomic and amino acid properties], Table 3SM [summary of input and output information for drugs and receptors] and Table 4SM [summary of input molecular parameters and predicted values for drugs and receptors]. This material is available free of charge via the Internet at http://pubs.acs.org.

MP800102C